Article:
Mendes, E.M.A.M. and Billings, S.A. (2001) An alternative solution to the model structure selection problem. IEEE Transactions on Systems Man and Cybernetics Part A: Systems and Humans, 31 (6). pp. 597-608. ISSN 1083-4427

https://doi.org/10.1109/3468.983416

An Alternative Solution to the Model Structure Selection Problem

Eduardo M. A. M. Mendes and Steve A. Billings

***Abstract*—An alternative solution to the model structure selection problem is introduced by conducting a forward search through the many possible candidate model terms initially and then performing an exhaustive all subset model selection on the resulting model. An example is included to demonstrate that this approach leads to dynamically valid nonlinear models.**

I. Introduction

IN THE FIELD of control engineering the task of system identification usually consists of determining a discrete linear or nonlinear mathematical model of a stochastic control system from the measurements of input and output signals. Although today's literature on parameter estimation schemes is vast, the final product of such schemes is often one single model which hopefully reproduces the system characteristics.

In traditional regression analysis, the problem of detecting an appropriate model based upon a subset of the original set of candidates consists of [1]

1) the computational algorithm used to provide information for the analysis;
2) the criterion used to analyze the candidates and select a subset;
3) the estimation of the coefficients in the final model.

In [2] it is argued that it is unlikely that there is a single best model but rather several equally good ones. This suggests that an evaluation of a fairly large number of models might be desirable.

When only dynamically validated models are of interest, criteria which incorporate such information during the process of identification should be used. In the context of neural network models, it has been shown that the minimization of the mean squared errors is likely to be a necessary condition for the model to reproduce the dynamical invariants of the original system, but it is definitely not sufficient [3]. These authors used such a statement to justify the utilization of *a posteriori* tests and to consequently verify the model's validity. In this paper similar tests will be introduced as a tool for selecting not only one single model but a family of models which can adequately reproduce the desired dynamical invariants of the original system. The primary objective is to show that the use of identification schemes which return many good models is worth exploring.

Many procedures which perform exhaustive search or related approaches have been investigated since the 1960s [e.g., [4]–[7]]. An excellent description of such methods and many others is given in [8]. Typically, these procedures consider all subsets of all sizes, and require that the number of data points is at least as great as the number of regressors (or terms). With these procedures it is possible to obtain the best model of each size. Different criteria have been devised in order to choose the optimal model and related good models. In [9] an algorithm was proposed for searching for all subsets of p or less variables out of k using the usual L_2-norm of deviations. Whereas Narula and Wellington [10] used the L_1-norm (i.e., minimization of the sum of absolute deviations to find the best models of p variables out of a k-variable trial model), several authors [1], [8], [11] warned users of such procedures about the computational demands imposed by generating all possible subsets. In [8] it is argued that an exhaustive search for all best fitting subsets is not feasible for trail models with more than 25 terms. Such a pitfall has been used to advocate the use of nonoptimal procedures such as *stepwise regression*, which often refers to an algorithm proposed by Efroymson [12], and other *forward* [13] and *backward* algorithms [See [8], for details]. These procedures are not without problems. In [14] it has been shown that *stepwise regression* does not always succeed in selecting the best subset of a determined size from a trial model when the chosen criterion is the minimization of the explained variance. The reason for this is that *stepwise regression* minimizes the increment to explained variance and not explained variance itself. Boyce *et al.* [14] argued that *optimal regression* (an exhaustive-like approach) and interdependence algorithms should be applied in place of *stepwise regression* and principal components analysis insofar as subset selection is concerned.

This paper uses both ideas, that is, *forward* and *exhaustive-like* searches to select models which can reproduce adequately the dynamical invariants of the system under investigation. Briefly, the procedure adopted here will be to conduct a forward search throughout the many possible candidate terms using the orthogonal estimator (OLS-ERR) [13] and then to use the resultant model to perform a similar procedure to *all subset selection* procedures. Section II includes some background material. In Section III a justification for performing a second search over the final OLS-ERR model is provided. Section IV describes the **second search procedure** and reviews the role of information criteria in selecting the trade-off between

Manuscript received May 26, 1998; revised June 18, 2001. E.M.A.M. Mendes acknowledges the support of CNPq under Grants 301313/96-2 and 201557/91-6. S. A. Billings acknowledges support from EPSRC. This paper was recommended by Associate Editor E. Page, Jr.

E. M. A. M. Mendes is with Departamento de Eletricidade, Fundação de Ensino Superior de São João Del Rei, São João Del Rei, Brazil (e-mail: emendes@funrei.br; emendes@mail.la.utexas.edu).

S. A. Billings is with the Department of Automatic Control and Systems Engineering, University of Sheffield, Sheffield, U.K. (e-mail: s.billings@sheffield.ac.uk).

Publisher Item Identifier S 1083-4427(01)09756-9.

complexity and fitting accuracy. An example is provided to demonstrate the benefits of the new approach. The main points of the paper are summarized in Section V.

II. Background Material

This section reviews the concepts used later in the paper.

A. NARMAX Approach

The nonlinear difference equation model, known as the Non-Linear Auto Regressive Moving Average with eXogenous inputs (NARMAX) model [15] provides a unified representation for a wide class of nonlinear systems. Leontaritis and Billings [15] showed that several well-known models such as the Hammerstein, Wiener and bilinear models are special cases of the NARMAX model [16]. This model can represented as follows:

$$\begin{aligned} y(k) = \mathtt{F}\,(&y(k-1), \ldots, y(k-n_y), u(k-d), \ldots, \\ &u(k-d-n_u), e(k-1), \ldots, e(k-n_e)) + e(k) \end{aligned} \tag{1}$$

where

$y(k)$, $u(k)$ and $e(k)$	output, input and noise, respectively;
n_y, n_u and n_e	corresponding maximum lags;
$\{e(t)\}$	accounts for possible noise, uncertainties, unmodeled dynamics;
$d \in \mathbb{Z}^+$	delay;
$\mathtt{F}$	some nonlinear function, the form of which is usually unknown.

Equation (1) is typically used in identification procedures since $e(k)$ is unknown in practice. Insofar as such procedures are concerned, the inclusion of monomials in $e(k)$ is mainly to avoid the bias in the parameters.

Because $\mathtt{F}$ can assume a variety of forms, the identification of nonlinear systems becomes a much more difficult task than the linear counter-part, where the major difficulty is to determine the system order. The uniqueness of the representation was addressed in [17] where the NARMAX model was referred to as the recursive representation of a system.

For many real sampled nonlinear systems, the exact NARMAX models, described by the function $\mathtt{F}$ in (1), are very difficult to determine. Therefore, it is often necessary to approximate $\mathtt{F}$ by some function. Polynomial NARMAX models have been shown to be a very good choice [18]. The theoretical justification for using polynomial NARMAX models to represent nonlinear systems have been given in [19]. Being linear-in-the-parameters, the polynomial models can readily be estimated using linear least squares methods as can be seen in Section II-B.

B. Orthogonal Least-Squares Estimator With Structure Detection (OLS-ERR)

In this section the orthogonal least-squares estimator with structure detection proposed by [20] is reviewed. In order to provide a practical application of such an estimator, all explanations will be based upon the estimation of polynomial NARMAX models. To this end, consider the polynomial NARMAX model based upon (1) in the following form:

$$y(k) = \Psi_{yu}^{\mathrm{T}}(k-1)\Theta_{yu} + \Psi_{yue}^{\mathrm{T}}(k-1)\Theta_{yue} + e(k) \tag{2}$$

where $\Psi_{yu}^{\mathrm{T}}(k-1)$ includes a constant and all the output and input terms as well as all combinations up to degree ℓ and time $k-1$. These terms will henceforth be referred to as *process terms*. The vector Θ_{yu} are the parameters of such terms. The matrix $\Psi_{yue}^{\mathrm{T}}(k-1)$ and the vector Θ_{yue} are defined likewise. $\Psi_{yue}^{\mathrm{T}}(k-1)$ will be referred to as *noise terms*.

Unfortunately (2) is not suitable for estimating the parameters of a polynomial NARMAX model because the noise terms are not known. However the noise sequence $e(k)$ can be estimated interactively as

$$\xi(k) = y(k) - \hat{y}(k|\hat{\Theta}) \tag{3}$$

where $\xi(k)$ is the residual at time k and $\hat{y}(k|\hat{\Theta})$, the prediction of $y(k)$, can be written as

$$\hat{y}(k|\hat{\Theta}) = \Psi_{yu}^{\mathrm{T}}(k-1)\hat{\Theta}_{yu} + \Psi_{yu\xi}^{\mathrm{T}}(k-1)\hat{\Theta}_{yu\xi}. \tag{4}$$

Finally, substituting (4) into (3) and rearranging the order yields

$$y(k) = \Psi^{\mathrm{T}}(k-1)\hat{\Theta} + \xi(k) \tag{5}$$

where $\Psi^{\mathrm{T}}(k-1) = \left[\Psi_{yu}^{\mathrm{T}}(k-1)\ \Psi_{yu\xi}^{\mathrm{T}}(k-1)\right]$ and $\hat{\Theta}^{\mathrm{T}} = \left[\hat{\Theta}_{yu}^{\mathrm{T}}\ \hat{\Theta}_{yu\xi}^{\mathrm{T}}\right]^{\mathrm{T}}$. Equation (5) clearly belongs to the linear regression model

$$y(k) = \sum_{i=1}^{n_\theta} p_i(k)\theta_i + \xi(k), \quad k = 1, \ldots, N \tag{6}$$

where

N	data length;
$p_i(t)$	column-vectors which represent *process* and *noise terms*;
n_θ	number of distinct such column-vectors;
θ_i	unknown parameters to be estimated.

n_θ is the summation of n_p process terms and n_n noise terms. The extended least squares algorithm [21] consists of estimating the *process terms* first and then using (3) to obtain the residual sequence. Once the residuals are calculated, the *noise terms* are incorporated into the matrix $\Psi^{\mathrm{T}}(k-1)$ and a new set of parameters Θ is estimated. This process is repeated until the residual sequence converges or a predetermined number of iterations is achieved.

Briefly, the basic principle of the orthogonal estimator proposed by Billings *et al.* [20] is to replace the original set of regression vectors by orthogonal vectors. The parameters associated with new vectors are such that the contribution of each vector can be measured independently of the rest of the vectors and these can then be related back to original model terms and variables. This simple result makes possible the selection of relevant terms in a polynomial model and unlike principal components and other similar methods the final model is expressed in terms of the original system variables.

In the orthogonal estimator the parameter estimation is performed for a linear-in-the-parameters model which is closely related to (6) and which can be represented as

$$\begin{aligned} y(k) &= \sum_{i=1}^{n_\theta} w_i(k) g_i + \xi(k), \quad k = 1, \dots, N \\ y &= [w_1 \dots w_{n_\theta}] G + \xi = \mathbf{W}\, G + \xi \end{aligned} \tag{7}$$

where the orthogonal vectors w_i and the parameters g_i are constructed from (6). The original parameters θ_i of the model in (6) can be calculated from the $\{g_i\}_{i=1}^{n_\theta}$.

As stated before, a great advantage of the orthogonal estimator is the possibility of selecting the relevant vectors (terms) as a by-product. To demonstrate this, consider again the orthogonal regression (7). In doing so, it is assumed that the orthogonal property $w_i^{\mathrm{T}} w_j = 0$ for $i \neq j$ holds. Therefore, if (7) is multiplied by itself and the time average is taken, the following equation can be derived:

$$\frac{1}{N} y^{\mathrm{T}} y = \frac{1}{N} \sum_{i=1}^{n_\theta} g_i^2 w_i^{\mathrm{T}} w_i + \frac{1}{N} \boldsymbol{\xi}^{\mathrm{T}} \boldsymbol{\xi}. \tag{8}$$

The output variance $y^{\mathrm{T}} y/N$ consists of two terms. The first term $\sum_{i=1}^{n_\theta} g_i^2 w_i^{\mathrm{T}} w_i/N$ is the part of the output variance explained by the regressors whereas the second term $\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\xi}/N$ accounts for the unexplained variance. Owing to the orthogonal estimator, the increment toward the overall output variance of each regressor (term or vector) can be computed independently as $g_i^2 w_i^{\mathrm{T}} w_i$. Expressing this quantity as a fraction of the overall output variance yields the *error reduction error* (ERR)

$$[\mathrm{ERR}]_i = \frac{g_i^2 w_i^{\mathrm{T}} w_i}{y^{\mathrm{T}} y}, \quad 1 \leq i \leq n_\theta. \tag{9}$$

ERR can be used as a simple and effective means of selecting the most relevant regressors in a forward-regression manner. Therefore, ERR imposes a hierarchy of terms according to their contribution toward the overall output variance.

C. Term Clustering

The deterministic part of a NARMAX model, that is, a NARX model, can be expanded as the summation of terms with degrees of nonlinearity in the range $1 \leq m \leq \ell$. Each mth-order term can contain a pth-order factor in $y(k - n_i)$ and a $(m - p)$th-order factor in $u(k - n_i)$ and is multiplied by a coefficient $c_{p,m-p}(n_1, \dots, n_m)$ as follows [22]:

$$\begin{aligned} y(k) = \sum_{m=0}^{\ell} \sum_{p=0}^{m} \sum_{n_1, n_m}^{n_y, n_u} & c_{p,m-p}(n_1, \dots, n_m) \\ & \times \prod_{i=1}^{p} y(k - n_i) \prod_{i=p+1}^{m} u(k - n_i) \end{aligned} \tag{10}$$

where $\sum_{n_1,n_m}^{n_y,n_u} \equiv \sum_{n_1=1}^{n_y} \cdots \sum_{n_m=1}^{n_u}$ and the upper limit is n_y if the summation refers to factors in $y(k - n_i)$ or n_u for factors in $u(k - n_i)$. In discrete models estimated from data generated from nonlinear continuous systems, the term coefficients depend on the sampling time T_s and should therefore be strictly represented as $c_{p,m-p}(T_s, n_1, \dots, n_m)$. However, for the sake of brevity, the argument T_s is dropped.

If the window of length $(n_y - 1) \times T_s$, defined by the model, is sufficiently short such that

$$\begin{aligned} y(k-1) \approx & y(k-2) \approx \cdots \approx y(k - n_y) \\ u(k-1) \approx & u(k-2) \approx \cdots \approx u(k - n_u) \end{aligned} \tag{11}$$

then (10) can be rewritten as

$$\begin{aligned} y(k) \approx \sum_{n_1, n_m}^{n_y, n_u} & c_{p,m-p}(n_1, \dots, n_m) \\ & \times \sum_{m=0}^{\ell} \sum_{p=0}^{m} y(k-1)^p u(k-1)^{m-p}. \end{aligned} \tag{12}$$

Definition II.1 [23]: The constants $\sum_{n_1,n_m}^{n_y,n_u} c_{p,m-p}$ $(n_1, \dots, n_m)$ in (12) are the coefficients of the *term clusters* $\Omega_{y^p u^{m-p}}$, which contain terms of the form $y(k-i)^p u(k-j)^{m-p}$ for $m = 0, \dots, \ell$ and $p = 0, \dots, m$. Such coefficients are called *cluster coefficients* and are represented as $\Sigma_{y^p u^{m-p}}$. □

Clearly, the set of candidate terms for a NARX model is the union of all possible clusters up to degree ℓ.

D. Fixed Points

The fixed points of a map are defined as those points for which $y(k) = y(k+i)$, $i \in \mathbb{Z}$ and usually constitute the starting point in the analysis of nonlinear systems [24].

Usually the fixed points are computed for the autonomous version of the system under investigation. If the original polynomial is nonautonomous, then set $u(k - i) = 0$, $i = 0, 1, \dots$ so that the only remaining terms are those involving the output. The resultant equation (or model) can be considered as an autonomous polynomial and can therefore be used for calculating the fixed points. All the possible clusters of an autonomous polynomial with degree of nonlinearity ℓ are $\Omega_0 = \mathrm{constant}$, Ω_y, $\Omega_{y^2}, \dots,$ Ω_{y^ℓ}.

Based upon this definition and using the cluster coefficients, the fixed points of an autonomous polynomial with degree of nonlinearity ℓ can be calculated by finding the roots of the following "clustered polynomial"

$$\Sigma_{y^\ell} y(k)^\ell + \cdots + \Sigma_{y^2} y(k)^2 + (\Sigma_y - 1) y(k) + \Sigma_0 = 0, \tag{13}$$

where $\Sigma_0 = c_{0,0}$ is a constant. From (13) it can be seen that an autonomous polynomial with degree of nonlinearity ℓ will have ℓ fixed points if $\Sigma_{y^\ell} \neq 0$. It should be pointed out that the fixed points are important in the model structure problem [25].

E. Correlation Tests

In the theory of linear systems, the usual statistical approach to validating identified linear models consists of computing the autocorrelation function of the residuals and the cross-correlation function between the residuals and the input [26].

It has been shown that acceptable predictions over different data sets are produced only if the model is unbiased. If the model structure and the estimated parameters are correct then the prediction error sequence $\xi(k)$ should be unpredictable from all linear and nonlinear combinations of past inputs and outputs and this will hold if and only if the following conditions are satisfied [27]:

$$\begin{aligned}\Phi_{\xi\xi}(\tau) =&E[\xi(k-\tau)\xi(k)] = \delta(\tau)\\ \Phi_{u\xi}(\tau) =&E[u(k-\tau)\xi(k)] = 0, \quad \forall\, \tau\\ \Phi_{u^{2'}\xi}(\tau) =&E\left[\left(u^2(k-\tau)-\overline{u^2(k)}\right)\xi(k)\right] = 0, \quad \forall\, \tau\\ \Phi_{u^{2'}\xi^2}(\tau) =&E\left[\left(u^2(k-\tau)-\overline{u^2(k)}\right)\xi^2(k)\right] = 0, \quad \forall\, \tau\\ \Phi_{(\xi)(\xi u)}(\tau) =&E[\xi(k)\xi(k-1-\tau)u(k-1-\tau)] = 0, \;\; \tau \geq 0\end{aligned} \tag{14}$$

where $\delta(\tau)$ is the *Kronecker delta*. The overbar indicates mean value and $E[\bullet]$ denotes the mathematical expectation. When no input is available, that is, the data are a time series, the following correlations test functions should be used [28]:

$$\begin{aligned}\Phi_{\xi'\xi'}(\tau) =&E\left[\left(\xi(k)-\overline{\xi(k)}\right)\left(\xi(k-\tau)-\overline{\xi(k)}\right)\right] = \delta(\tau)\\ \Phi_{\xi'(\xi^{2'})}(\tau) =&E\left[\left(\xi(k)-\overline{\xi(k)}\right)\left(\xi^2(k-\tau)-\overline{\xi^2(k)}\right)\right] = 0\\ \Phi_{\xi^{2'}\xi^{2'}}(\tau) =&E\left[\left(\xi^2(k)-\overline{\xi^2(k)}\right)\left(\xi^2(k-\tau)-\overline{\xi^2(k)}\right)\right]\\ =&\delta(\tau).\end{aligned} \tag{15}$$

Recently two new correlation functions were introduced in [29] as a solution to increase the discriminatory power of the existent correlation tests (14). These two correlation functions defined in terms of delayed outputs are

$$\begin{aligned}\Phi_{(y\xi)'\xi^{2'}}(\tau) =&E\left[\left(y\xi(k)-\overline{y\xi(k)}\right)\left(\xi^2(k-\tau)-\overline{\xi^2(k)}\right)\right]\\ =&k_\xi\delta(\tau),\\ \Phi_{(y\xi)(u^{2'})}(\tau) =&E\left[\left(y\xi(k)-\overline{y\xi(k)}\right)\left(u^2(k-\tau)-\overline{u^2(k)}\right)\right]\\ =&k_\xi\delta(\tau)\end{aligned} \tag{16}$$

where the constant k_ξ is defined in [29].

The underlying rationale of the correlation tests (14)–(16) is that for a model to be statistically valid, there should be no predictable terms in the residuals. However, in practice only a finite data length will be available. This implies that confidence bands should be used to reveal if the correlation between variables is significant or not. For large N the 95% confidence bands are approximately $\pm 1.96/\sqrt{N}$ and any significant correlation will be indicated by one or more points of the function lying outside these bands.

III. Justification for Performing a Second Term Search

There is a considerable literature on the subject of selection of the "best" subset out of a trial model. As pointed out in the introduction, the majority of procedures used in practice do not take into consideration dynamical characteristics of the system. Instead they rely on statistical measures. In this section an attempt to justify a second term search over a predetermined trial model is made. To this end, it is necessary to introduce some important concepts.

The structure of identified models can be characterized in $\mathbb{R} \times \mathbb{N}^2$ space by a point of coordinates $(T_s, n_p, \max(n_y, n_u) = m)$ where T_s is the sampling time, n_p is the number of terms allowed in the final model and $\max(n_y, n_u) = m$ is the maximum value between the maximum values of output and input lag, respectively. m is the usual embedding dimension extensively used in the literature. The space defined as above is denoted *model structure space* (MSS) [30]. In this reference a subregion $\mathcal{Q}$ of the MSS is defined where the best estimated models are located in accordance with the free parameters. It has been demonstrated that both N, the number of data points, and the noise terms are extremely important and can determine whether an estimated model reproduces the desired dynamical characteristics or not. For instance, it can be shown that a model with 18 process terms and 20 linear noise terms reproduces the double scroll attractor of Chua's circuit [31]. But if no noise terms were included in the final model, this very same model (structure) exhibits a completely different motion. In the light of these simple but striking results, the subregion $\mathcal{Q}$ should always be defined in terms not only of $(T_s, n_p, \max(n_y, n_u))$ but also of N, the noise variance and noise terms.

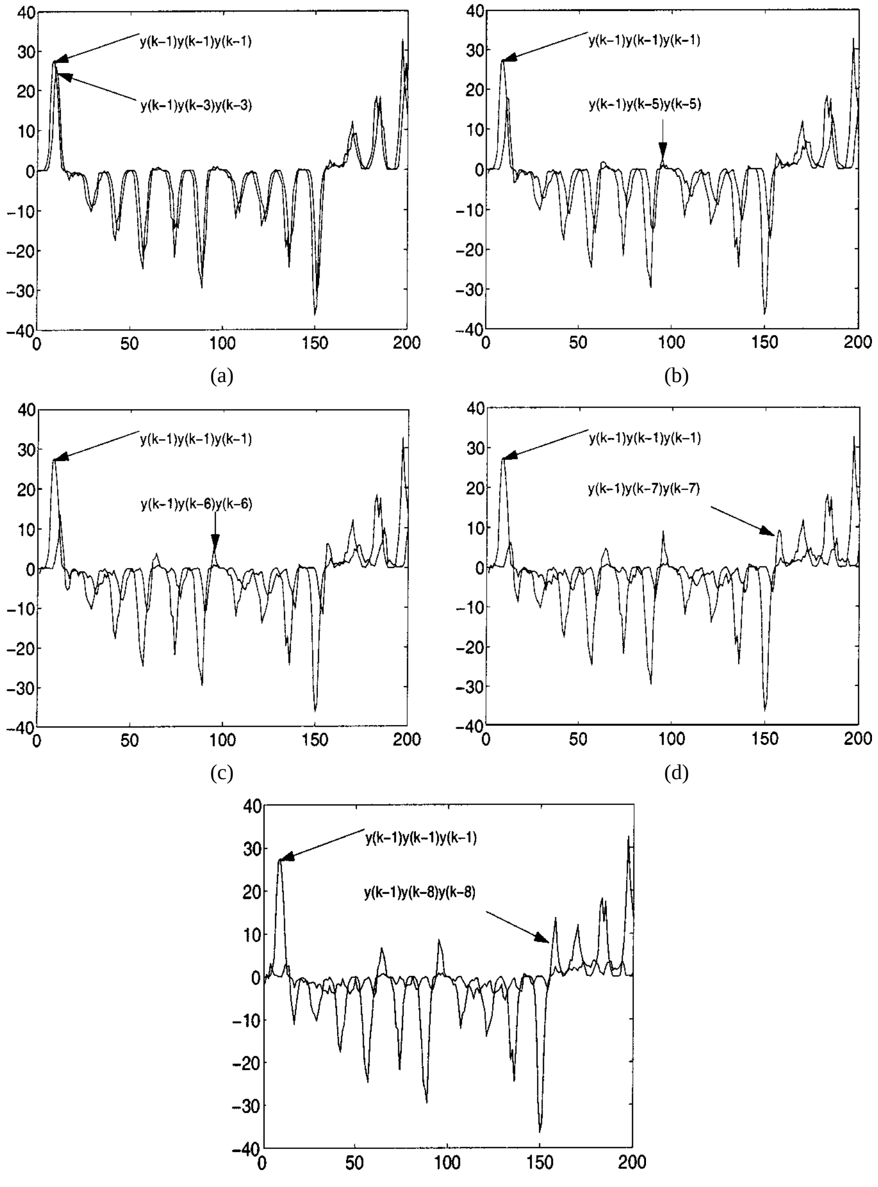
As for the influence of the sampling time on the identification of valid models, the discussion in [32] provides clear evidence that the cautious use of higher degree of nonlinearity can increase drastically the number of good models.[1] Moreover, such models can have n_p greater than the values specified in [30]. Therefore, it is conjectured that the number of valid models in $\mathcal{Q}(T_s, n_p, n_y, N, \sigma_\xi,$ noise terms) is much greater than previously considered. This constitutes the basis of the argument that a second search over a predetermined identified model is worth considering. These results appear to contradict a recent numerical experiment conducted in [33]. In this paper 10^6 models were estimated from data generated by the Lorenz [34] and Rössler [35] equations. They reported that 99% of the estimated models were unstable and that only about 0.04% of the model exhibit some chaotic motions but not which were necessarily valid. The reason for these low figures appears to be the estimation of models with 60 coefficients and no noise terms.

The next example will be used to introduce the concept of *model families*. Also it will be shown that if the correct number of noise terms is chosen rather large models can be identified.

Example 1: Consider the set of normalized equations (x, y, z) of Chua's circuit [31]. The equations of motion of such a circuit were used to generate data for identification purposes. The resultant data of the z-coordinate sampled at $T_s = 0.15$ were then corrupted by white noise so that the signal-noise ratio was approximately 42 dB. The number of data points considered in this analysis was $N = 1801$.

To demonstrate that even for a large number of terms a reasonably good model can be estimated, certain conditions must be satisfied: 1) $n_y \leq n_{y.\max}$ where $n_{y.\max}$ is the upper bound for the number of degrees of freedom (DOF) required to describe the system dynamics and 2) no *spurious* clusters should be in the final model. To avoid the presence of such clusters, it has been shown that the fixed points and consequently the related clusters can be estimated directly from the data using the procedure of Glover and Mees [36] as described in [37]. Once the true clusters are determined, the next step is to then roughly estimate the value of $n_{y.\max}$. This can be done visually for

[1] This seems to be true even for rather coarse values of T_s. However, it is worth mentioning that if T_s is increased beyond a certain value defined by the sampling theorem (two times the Nyquist frequency), the information about the system dynamics will be lost and cannot be retrieved by estimation procedures.



(a)

(b)

(c)

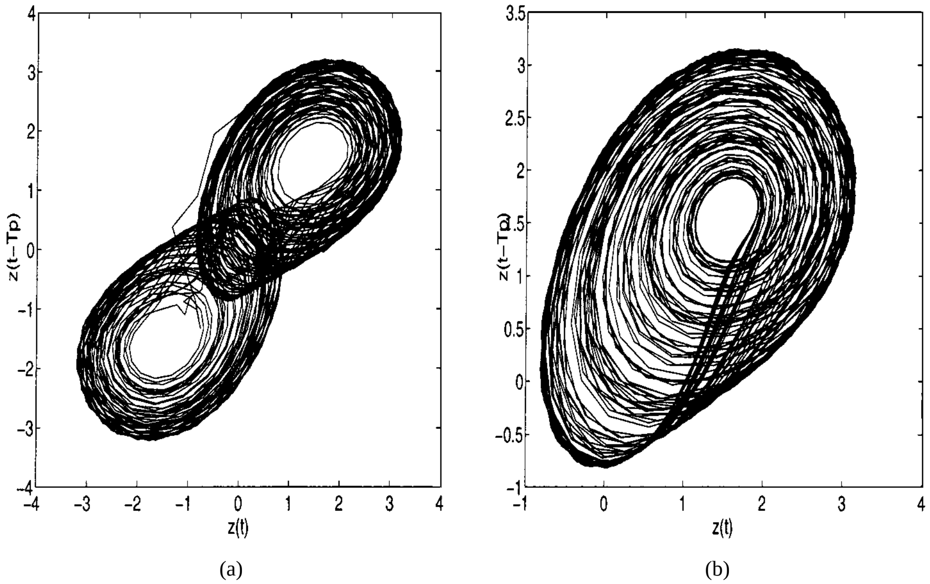
(d)

(e)

Fig. 1. Estimation of $n_{y.\max}$ using data from the z-coordinate of Chua's circuit corrupted by white noise, showing plots of the time series generated by different terms of the cluster Ω_{y^3}. The vertical axis is the magnitude of the series whereas the horizontal axis is the number of data points. Note that with only 200 points the maximum lag $n_{y.\max}$ can be estimated.

data originated from sampling a continuous-time system. Before showing this procedure, it is worth mentioning that in [30] it is suggested that the lower bound provided by Takens' theorem [38] is often larger than necessary [39] and could therefore be used as an estimate of $n_{y.\max}$. The minimum number of degrees of freedom $n_{y.\min}$ can be obtained, for instance, from estimates of the fractal dimension D_0. $n_{y.\min}$ is always larger than D_0 [40]. Abarbanel *et al.* have recently devised methods for detecting the minimum embedding dimension [41]. Such a dimension can also be estimated as a by-product of Savit and Green's procedure [42].

The procedure used to obtain a rough estimate of $n_{y.\max}$ is based upon the fact that $y(k-1) \approx \cdots \approx y(k-n_{y.\max})$ for appropriate values of the sampling time T_s. Since the identified models are usually nonlinear, it is conjectured that a similar relation involving nonlinear terms is also valid. For values of n_y greater than of $n_{y.\max}$ this relation is no longer observed. These ideas are better illustrated in Fig. 1(a)–(e). Clearly, it can be seen that terms such as $y(k-1)y(k-8)^2$ no longer describe a similar trajectory as that of terms like $y(k-1)^3$ or $y(k-1)y(k-3)^2$. This pattern has been observed not only for Chua's circuit but also for other systems such as Duffing–Ueda [43], Duffing–Holmes [44], etc. From these figures $n_{y.\max} = 6$ is a rough estimate of the maximum number of degree of freedom valid for identification purposes because for lags higher than six the cubic terms no longer have similar trajectories to those of terms of lags lower than six. Values beyond this upper bound appear not to be recommendable for identification of dynamically valid models. In [45], it is stated that models with $n_y = 6$ do not reproduce the chaotic effects as accurately as models with $3 < n_y < 6$. It is believed that, although it might be possible to identify models with $n_y \geq 6$, they are bound



(a)

(b)

Fig. 2. (a) Double scroll attractor obtained from data generated by the full **trial model** with one noise term. (b) Spiral-like attractor obtained from data generated by the full **trial model** with five noise terms. $Tp = 3$ was used.

to produce spurious dynamics because of the presence of extra Lyapunov exponents, as discussed in [37], which causes overfitting due to the introduction of higher lag terms. These terms also tend to explain the noise in the data.

Taking into consideration all the previous discussion, a **trial model** with $n_y = 5$ and $\ell = 3$ was chosen. Only linear and cubic terms were allowed due to the attractor's symmetry detected by Glover and Mees' procedure. The total number of (process) terms of the resultant **trial model** is 40 (plus noise terms), that is, all terms were then forced in the regression.

It is important to note that the noise terms play an important role in the determination of the validity of an identified model. For instance, if only a single noise term was included in the **trial model** reproduces the dynamical invariants fairly well despite the excessive number of terms. Fig. 2(a) shows the double scroll attractor reproduced from the identified **trial model**. Note however that the noise term $\xi(k-1)$ is not enough to "bleach" the data and therefore the residuals are correlated [see especially Fig. 3(a)]. As more noise terms are included in the model, it can be observed that 1) all correlation functions lie inside the respective confidence bands and 2) several different chaotic motions are registered indicating extreme sensitivity to parameter estimation.[2] Fig. 2(b) shows a chaotic motion which resembles the spiral attractor. Table I displays the Lyapunov spectrum for the two models. Note that these models are *dimension overparametrized* which can cause a large variety of dynamical behaviors not exhibited by the original system.

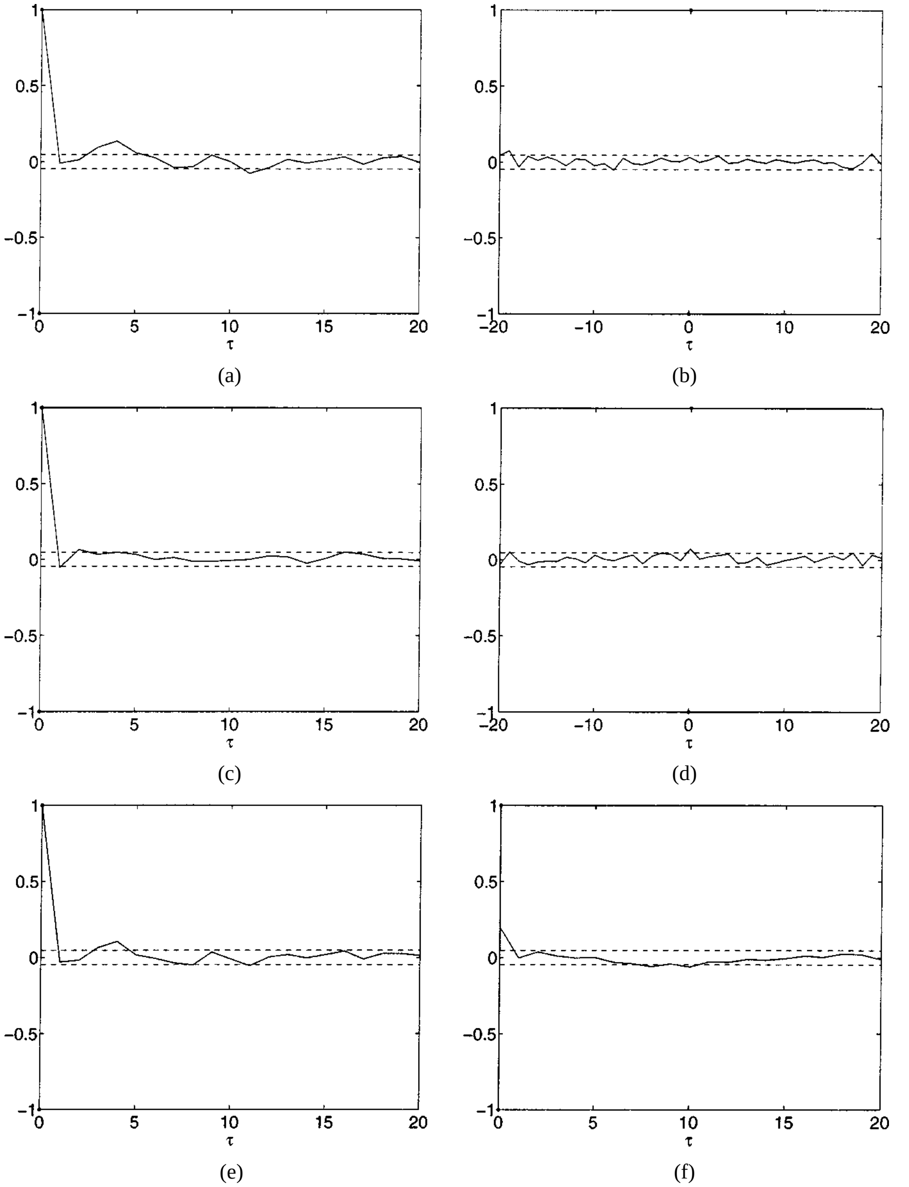
The fixed points of the **trial model** are $(0, 1.564, -1.564)$ which compares quite well with the fixed points calculated from the equations of Chua's circuit $(0, 1.5, -1.5)$. When compared to the previous model for the z-coordinate of Chua's circuit, it can be noticed that the location of the fixed points remain almost constant regardless of the number of terms in the model. Throughout it is assumed that only the effective clusters are present in the trial model. It seems that the orthogonal estimator is trying to accommodate the model coefficients in order to preserve the location of the fixed points. This becomes more difficult when a large number of terms is involved in the calculation due to ill-conditioning of the numerical solution. The primary problem is not only to obtain a similar location of the fixed points but also stable models. This could be rather difficult to achieve since it has been shown that small variations in parameters lead to completely different dynamical behaviors. The same conclusion has been stated in [47] for high dimensional systems. Therefore, variations of the combination of linear and cubic terms could lead to the identification of valid models which were not selected by the OLS-ERR procedure. This statement justifies a second search over a model previously identified using such a procedure. Apart from the computational demands imposed by an exhaustive search, it will be shown that a sub-optimal procedure can result in better models in cases where the OLS-ERR procedure fails to detect a few good models.

The set of models which exhibits similar characteristics will be denoted as the *model family*. In the case of models identified from the z-coordinate of Chua's equations the *model family* is all models with $n_y \leq n_{y\,\max} = 6$ and linear and cubic terms. This constitutes a *model family* since the fixed points calculated from the model equation for each member are similar and moreover placed near the original fixed points $(0, -1.5, 1.5)$. Several members may exhibit very similar dynamical characteristics and could, therefore, be considered as "optimal" models. This will be illustrated shortly. □

IV. Second Search Procedure

One method of selecting subsets (or terms) in regression problems is *stepwise regression* [12] or closely related procedures, called forward selection and backward elimination. The deficiencies of these one-variable-at-time methods have been discussed in [1] and [11]. Hocking [11] points out that "the primary problems are that 1) the procedure implies an order of

[2] Theiler and Eubank [46] have shown that the bleaching is more effective when the order of the moving average part (n_e) is large.



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 3. Correlation tests of the residuals $\xi(k)$ originated from the identification of the full **trial model**. (a) $\Phi_{\xi'\xi'}(\tau)$. (b) $\Phi_{\xi'\xi^{2'}}(\tau)$. (c) $\Phi_{\xi^{2'}\xi^{2'}}(\tau)$. (d) $\Phi_{\xi^{2'}\xi^{3'}}(\tau)$. (e) $\Phi_{\xi^{3'}\xi^{3'}}(\tau)$. (f) $\Phi_{(y\xi)'\xi^{2'}}(\tau)$.

TABLE I
COMPARISON BETWEEN THE ORIGINAL LYAPUNOV SPECTRUM OF CHUA'S CIRCUIT AND THAT OF THE FULL TRIAL MODEL

	Double Scroll		Spiral	
	Original	Estimated[a]	Original	Estimated[b]
True	0.23	0.2108	0.09	0.1089
	0	-0.0000626	0	-0.0003333
	-1.78	-2.363	-1.49	-2.187
Spurious	-	-2.538	-	-2.26
	-	-2.984	-	-3.079

[a] One single noise term
[b] Five linear noise terms

importance to the variables, an order that may be misleading or confusing, and 2) in the case of early termination, the procedure fails to detect important variables." Although these problems occurred in such procedures an alternative method using the orthogonal estimator with structure selection [13], [48]–[50] has been shown to be very efficient in most cases. A similar procedure can be found in [51]. In this case the authors incorporate statistical tests in order to avoid the inclusion of unwanted terms.

An obvious alternative to circumvent the aforementioned problems is to evaluate all possible subsets, that is, to consider all $2^{n_p} - 1$ choices of equations involving $1, 2, \ldots, n_p$ variables from the original set of n_p variables. The so-called all subset selection procedures are methods of selecting variables that minimize a determined criterion such as the minimum residual sum of squares using some ingenious numerical techniques. These methods have been discussed by several authors and amount to selecting the "best" model from a particular set of variables. For an old but useful review of such methods see [1] and [11].

All subset selection procedures usually demand lots of computation time. Hocking and Leslie [52] developed a procedure for determining the subset of each size with minimum residual sum of squares without evaluating all possible regressions. Several other procedures have been published using the same kernel idea (see for example [6] and [14]). When the number of variables to be selected is very large, the all subset procedures become nonviable. This is often the case when trying to fit a nonlinear model to data.

In this section a method denoted henceforth as the *second search procedure* which incorporates the advantage of forward and all subset methods is proposed. Three examples are provided to illustrate the benefits of applying such a procedure. Numerous interesting issues concerning identification are also discussed and solutions are reported. But first, a brief review of some available information criteria is made.

A. Information Criteria

One of the issues when different models structures have been identified is how to compare them and decide which one (or ones) is the best. In statistical terms, the most natural and straightforward method is to evaluate prediction error variances of the different models over new data, that is, data not used for structure selection and parameter estimation. This procedure is called cross-validation [53]. In this method, if a model predicts better over the new data set then it should be considered as the best model.

Ljung and Glad [54] argued that cross-validation always requires a fresh amount of data for model comparison, and therefore not all available information is used for identification. When the model comparison has to be made over the same data used for estimation, a simple prediction error criterion cannot be used. The reason for this is that a larger model always gives a lower variance of prediction errors. Therefore, a trade-off between the number of terms in a model and its capacity to reduce the variance of the prediction errors should be sought. In the field of statistics, this trade-off is achieved by different methods which are, in general, based upon information theoretical principles [54]. These methods have similar characteristics, that is, they consist of a determined function which increases with the number of terms (n_p) and decreases with the number of data points (N). Minimizing this function with respect to n_p penalizes models which contain an excessive number of parameters or models with large variance of the prediction errors.

The most well-known methods available in the literature are

1) *Akaike's information criterion* (AIC) [55]

$$\mathrm{AIC}(\alpha) = N\log_e\left[\sigma_\xi^2(n_p)\right] + \alpha n_p, \quad \alpha > 0. \tag{17}$$

2) *Final prediction error* (FPE) [55]

$$\mathrm{FPE} = N\log_e\left[\sigma_\xi^2(n_p)\right] + N\log_e\frac{N+n_p}{N-n_p}. \tag{18}$$

3) Khundrin's *law of iterated logarithm criterion* (LILC) [56]

$$\mathrm{LILC} = N\log_e\left[\sigma_\xi^2(n_p)\right] + 2n_p\log_e\log_e N. \tag{19}$$

4) *Bayesian information criterion* (BIC) or Schwarz criterion [57]

$$\mathrm{BIC} = N\log_e\left[\sigma_\xi^2(n_p)\right] + n_p\log_e N. \tag{20}$$

5) Rissanen's *minimal description length* (MDL) [58]

$$\mathrm{MDL} = \frac{1}{2}\left(N\log_e\left[\sigma_\xi^2(n_p)\right] + n_p\log_e N\right). \tag{21}$$

In the criteria described above $\sigma_\xi^2(n_p)$ is the variance of the residuals associated to the n_p-term model. Some of the criteria listed above and others available in the literature are just modifications of the AIC criterion [e.g., [57], [59], [60]]. In [61] asymptotic comparisons of some of these criteria has been made. Stone has shown that there is an asymptotic equivalence between cross-validation and AIC. Other criteria such as Mallow's C_p [62], PRESS [63] and *model entropy* [64] can also be used as a tool for selecting a model which is a compromise between goodness of fit and complexity. The trade-off model structure is indicated by the value of n_p for which the chosen criterion reached a minimum value. For a review of some of these criteria refer to [65].

Note that the choice of stopping rule depends upon 1) the objectives and 2) the estimation method. In the case of estimation of nonlinear polynomial models, the estimation method used is the orthogonal estimator with structure detection. This estimator imposes an order by which the terms are selected. When calculating an information criterion, the user should be aware that the minimum value depends upon the chosen structure. Variations in this structure will lead to a different choice of number of terms.

The model selection problem using information criteria is still an active field of research. For instance, a new model selection criterion based on the Fisher information matrix was recently proposed in [66]

$$\mathrm{FIC}(n_p) = N\sigma_\xi^2(n_p) + \tilde{\sigma}_\xi^2\log_e\det\left(\sum_{i=1}^{n}\mathbf{P}_i\mathbf{P}_i^{\mathrm{T}}\right) \tag{22}$$

where $\sigma_\xi^2(n_p)$ and $\tilde{\sigma}_\xi^2$ are the variance of the residuals obtained from the process of identification of a n_p-term model and the full model, respectively.

The quantity $\left(\sum_{i=1}^{n}\mathbf{P}_i\mathbf{P}_i^{\mathrm{T}}\right)$ can be interpreted as the amount of information in the conditional Fisher matrix β defined as $\sigma^{-2}\left(\sum_{i=1}^{n}\mathbf{P}_i\mathbf{P}_i^{\mathrm{T}}\right)$. The main property of the Fisher Information Criteria (FIC) is that the conventional penalty term is replaced by a term that is proportional to the logarithm of the statistical information contained in a n_p-term model.

Despite the good characteristics, the FIC criterion demands heavy computations, when nonlinear models are concerned. For such models, the number of terms can well exceed thousands and therefore the estimation of $\tilde{\sigma}_\xi^2$ might lead to spurious results mainly due to ill-conditioned problems. Such problems also occurred when criteria such as Mallow's C_p [62] are used. It is conjectured that the usefulness of criteria which explicitly use the variance of the residuals of the full model is rather limited in nonlinear identification problems.

In the context of nonlinear polynomial models information criteria have been suggested by Kortmann and Unbehauen [67] as a means of determining a clear stopping rule for the number of terms. Whereas in [68] the Schwarz criterion is used as a stopping rule for radial basis function models, the information criteria can also be seen as the quantity to be minimized in the all subset selection methods.

In Example 1, the second search procedure is performed in order to minimize the difference between the first Lyapunov exponent estimated from the original system and the one calculated from the model's equation estimated directly from the data. The above criterion was adopted to access the validity of the models in terms of dynamical characteristics which cannot be done using the usual information criteria. It will be shown that several models can be identified using such a procedure.

Example 1: Consider again Chua's circuit equations used in Example 1 presented in Section III to demonstrate that when the model structure is correct even models with a rather large number of terms can reproduce the desired characteristics. A set of 1750 data points without noise contamination were considered to demonstrate the benefits of a second search over a pre-defined set of terms. In this work, this set of terms is always a model estimated using the orthogonal estimator with structure detection.[3] Typically, this model contains more terms than are necessary.

The experiment in this example was conducted as follows. Only models with $n_y = 5$, 10 linear noise terms $\left(\Psi_{\xi}^{\mathrm{T}}(k-1) = \xi(k-i),\ i = 1, 2, \ldots, 10\right)$ and a total number of terms less than 26 were used to define the subregion $\mathcal{Q}\ (T_s = 0.15, n_p \leq 25,\ n_y = 5,\ N = 1750,\ \sigma_{\xi} = 0,\ \text{noise terms} = 10\ \text{linear})$. This represents only a fraction of the valid models identified from the data generated by integrating Chua's equations. A small change in the number of noise terms can lead to valid models which were not considered as such when another noise configuration is adopted.

The first step is then to estimate models with increasing complexity using the OLS-ERR procedure. A set of 25 models with the number of terms varying from 1 to 25 were identified from the noise-free data of the z-coordinate of Chua's circuit. In [45, p. 858] it is suggested to use 20 linear terms in order to produce unbiased models. These authors stated that the residuals are white and zero-mean with variance $\sigma_{\xi}^2 = 0.925 \times 10^{-5}$. In the case of the noise-free data it is argued that the residuals are not white but simply modeling errors. If schemes such as the procedure discussed in [32] is used to integrate Chua's equations, these errors are the contribution of terms of degrees of nonlinearity higher than three and cannot, therefore, be considered as white noise. Fig. 4 shows the autocorrelation of the residuals obtained from the estimation of a model with 16 process terms and 20 linear noise terms (special case). Clearly the autocorrelation function shows some peaks lying outside of the confidence bands. Note that the bleaching effect [46] tends to obscure these peaks.

The dynamical invariants of the models estimated using OLS-ERR procedure were assessed by calculating the Lyapunov dimension and spectrum. These quantities are shown in Table II. The values of the Lyapunov dimension and spectrum for the estimated models demonstrates that only models with $n_p = 14$ and 16 can reproduce fairly well those of the original system. This result is a perfect agreement with those of [45].

The fixed points calculated directly from the equation of the aforementioned models are displayed in Table III. Note that all models with n_p from 9 to 24 exhibit the location of fixed points near the true value, that is, $(0, -1.5, 1.5)$ which signifies that these models could, in principle, reproduce the desired chaotic behavior. Furthermore the minimum number of terms appears to be $n_p = 9$, since all models with $n_p < 9$ have fixed points placed far from the correct location.

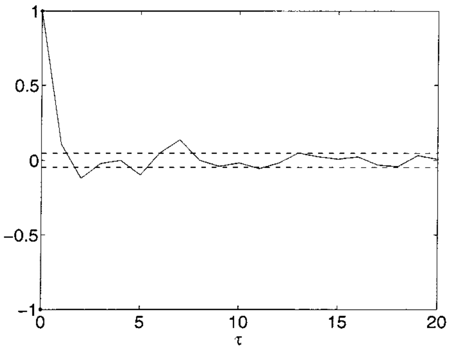


Fig. 4. Autocorrelation of the residuals obtained during the process of identification of models from data of the z-coordinate of Chua's circuit. In this particular case, the model contains 16 process terms and 20 linear noise terms. Variance of residuals is $\sigma_{\xi}^2 = 8.319 \times 10^{-6}$.

TABLE II
LYAPUNOV EXPONENTS FOR MODELS IDENTIFIED FROM NOISE-FREE DATA. A SINGLE TIME-SERIES OF THE z-COORDINATE OF CHUA'S CIRCUIT WAS CONSIDERED. ALTHOUGH THERE IS NO CONTAMINATION, TEN LINEAR NOISE TERMS WERE ALLOWED IN THE FINAL MODEL. OLS-ERR WAS USED TO SELECT THE STRUCTURE AND ESTIMATE THE PARAMETERS

n_p	Lyapunov Spectrum					D_L
	True Exponents			Spurious Exponents		
	λ_1	λ_2	λ_3	λ_4	λ_5	
TRUE	0.23	0.0	-1.78	-	-	2.13
≥ 17	-	-	-	-	-	Inf
16	0.2587	0.00054	-1.533	-3.297	-3.577	2.169
15	0.0664	-0.01908	-1.898	-4.303	-4.524	2.025
14	0.2655	-0.00602	-1.953	-3.547	-3.728	2.133
13	0.1854	0.00028	-0.6054	-0.6312	-3.691	2.307
12	-	-	-	-	-	Inf
11	-	-	-	-	-	Inf
10	-	-	-	-	-	Inf
9	0.02225	-0.06372	-0.1583	-0.182	-1.336	1.349
8	0.00212	-0.3836	-0.388	-1.143	-1.95	1.006
7	0.00349	0.00179	-0.01731	-1.462	-1.829	2.305
6	-	-	-	-	-	Inf
5	-0.1157	-0.116	-0.2364	-1.054	-1.054	0

Consider now a second search over the set of terms which comprises the model with $n_p = 25$ identified using the OLS-ERR procedure. The criterion adopted in this search was the minimization of the error between the first Lyapunov exponent of the original system and that of the estimated models. If the minimization were performed over all subsets generated from the 25-term long model, the parameters of $2^{25} - 1$ models would be estimated. Moreover, for each one of the estimated models, the first Lyapunov exponent was to be calculated which would certainly demand lots of computational time. In this example, rather than evaluating all these models, only a small fraction will be considered. The search starts from the estimated 25 term long model. All 25 models obtained by deleting one term at a time are evaluated and the model which exhibits the lowest error is kept for further search. This process

[3] Note that in Example 1, 1801 data points were considered. The choice of 1750 was done in order to compare the results presented in this present example with the ones in [45].

TABLE III
FIXED POINTS FOR MODELS ESTIMATED FROM THE NOISE-FREE DATA OF THE z-COORDINATE OF CHUA'S CIRCUIT. OLS-ERR WAS USED TO SELECT THE STRUCTURE AND ESTIMATE THE PARAMETERS

n_p	Fixed Points		
	FP_1	FP_2	FP_3
TRUE	0	-1.5	-1.5
24	0	1.512	-1.512
23	0	-1.512	1.512
22	0	1.508	-1.508
21	0	1.508	-1.508
20	0	1.513	-1.513
19	0	1.522	-1.522
18	0	-1.531	1.531
17	0	1.516	-1.516
16	0	1.51	-1.51
15	0	1.519	-1.519
14	0	1.525	-1.525
13	0	1.534	-1.534
12	0	1.578	-1.578
11	0	-1.562	1.562
10	0	1.558	-1.558
9	0	-1.574	1.574
8	0	0-1.099i	0+1.099i
7	0	0.6853	-0.6853
6	0	1.988	-1.988
5	0	-	-

is repeated until the minimum number of terms chosen *a priori* is reached.

The results of applying the above procedure are shown in Table IV. Clearly it can be seen that the number of valid (good) models has increased substantially compared to those of Table II. A wide range of models with n_p form 7 to 19 can reproduce the original dynamical invariants. This provides further support to the ideas discussed in the introduction of this paper.

Fig. 5 shows the double scroll attractor reconstructed from the identified model in (23). Despite the small number of terms, this model can reproduce fairly well the desired dynamical invariants.

$$\begin{aligned} z(k) = &+ 0.233\,63 \times 10^{+1} z(k-1) \\ &- 0.146\,53 \times 10^{+1} z(k-2) \\ &+ 0.195\,33 \times 10^{+0} z(k-5) \\ &- 0.452\,99 \times 10^{-1} z(k-1) z(k-1) z(k-2) \\ &+ 0.125\,76 \times 10^{+0} z(k-1) z(k-1) z(k-4) \\ &- 0.113\,07 \times 10^{+0} z(k-1) z(k-1) z(k-5) \\ &- 0.139\,36 \times 10^{-2} z(k-5) z(k-5) z(k-5) \\ &+ 0.680\,43 \times 10^{-2} z(k-1) z(k-2) z(k-5) \\ &+ \Psi_{\xi}^{\mathrm{T}}(k-1)\hat{\Theta}_{\xi} + \xi(k). \end{aligned} \tag{23}$$

The location of the fixed points for the models estimated using the second search procedure is displayed in Table V. It is interesting to note that the minimum number of terms required for a model to reproduce the original location of the fixed points is now 6. That is, the second search procedures seems to find better models with less terms. These results clearly demonstrate that the sub region $\mathcal{Q}$ consists of a much larger number of valid models than previously expected [30] and moreover that a second search is well worth the required computational demands.

TABLE IV
LYAPUNOV EXPONENTS FOR MODELS IDENTIFIED FROM NOISE-FREE DATA. A SINGLE TIME-SERIES OF THE z-COORDINATE OF CHUA'S CIRCUIT WAS CONSIDERED. ALTHOUGH THERE IS NO CONTAMINATION, TEN LINEAR NOISE TERMS WERE ALLOWED IN THE FINAL MODEL. SUBOPTIMAL PROCEDURE WAS USED TO SELECT THE STRUCTURE AND ESTIMATE THE PARAMETERS

n_p	Lyapunov Spectrum					D_L
	True Exponents			Spurious Exponents		
	λ_1	λ_2	λ_3	λ_4	λ_5	
TRUE	0.23	0.0	-1.78	-	-	2.13
≥ 20	-	-	-	-	-	Inf
19	0.2338	-0.01543	-1.374	-2.155	-3.709	2.159
18	0.2509	-0.00212	-1.746	-2.949	-3.625	2.142
17	0.2215	-0.0104	-1.272	-1.913	-3.538	2.166
16	0.277	-0.03315	-2.039	-3.013	-3.592	2.12
15	0.2757	-0.02215	-1.711	-2.549	-3.565	2.148
14	0.2838	-0.04172	-1.832	-2.271	-2.733	2.132
13	0.2839	-0.00055	-2.15	-3.62	-3.741	2.132
12	0.2478	-0.00005	-2.13	-3.703	-3.813	2.116
11	0.2653	-0.00039	-2.005	-3.752	-3.858	2.132
10	0.2376	-0.00140	-1.871	-3.776	-3.875	2.126
9	0.2243	-0.00088	-1.537	-3.736	-3.884	2.145
8	0.2363	-0.00049	-1.759	-3.698	-3.853	2.134
7	0.249	-0.00054	-1.517	-4.458	-4.6	2.164
6	0.0732	-0.00104	-1.109	-4.579	-4.707	2.065
5	0.0014	-0.36150	-0.3705	-4.925	-5.036	1.004

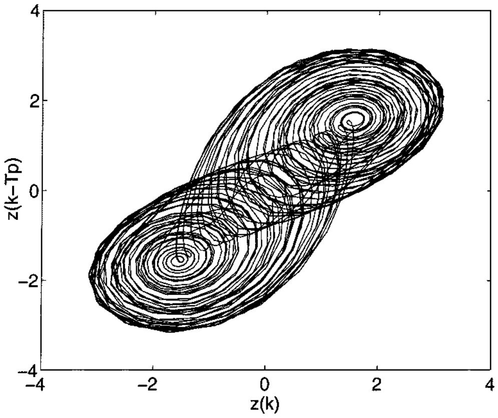


Fig. 5. Double scroll attractor reconstructed from the data generated by iterating (23). $Tp = 3$ was used.

Finally, to illustrate the results presented above in terms of information criteria the AIC criterion was calculated for the models estimated using both procedures, that is, OLS-ERR and the second search procedures. Fig. 6 shows the AIC curves for such models. Note the plateau which ranges from $n_p = 6$ to $n_p = 13$. These models have been shown to be able to reproduce fairly well the original dynamical invariants, however the residuals obtained during the process of estimation of such models exhibit higher variance than those obtained using the OLS-ERR procedure. This shows, not surprisingly, that lower residual variance does not signify better models. □

The second search procedure was also extended for identification of multivariable systems. An example of such a system is the nonlinear equations proposed by Rossler in [35]. When

TABLE V
FIXED POINTS FOR MODELS ESTIMATED FROM THE NOISE-FREE DATA OF THE z-COORDINATE OF CHUA'S CIRCUIT. SUBOPTIMAL PROCEDURE WAS USED TO SELECT THE STRUCTURE AND ESTIMATE THE PARAMETERS

n_p	Fixed Points		
TRUE	0	-1.5	-1.5
24	0	-1.512	1.512
23	0	-1.512	1.512
22	0	-1.508	1.508
21	0	-1.508	1.508
20	0	-1.513	1.513
19	0	-1.506	1.506
18	0	-1.506	1.506
17	0	-1.51	1.51
16	0	-1.518	1.518
15	0	-1.52	1.52
14	0	-1.524	1.524
13	0	-1.555	1.555
12	0	-1.558	1.558
11	0	-1.563	1.563
10	0	-1.563	1.563
9	0	-1.544	1.544
8	0	-1.562	1.562
7	0	-1.581	1.581
6	0	-1.563	1.563
5	0	-1.856	1.856

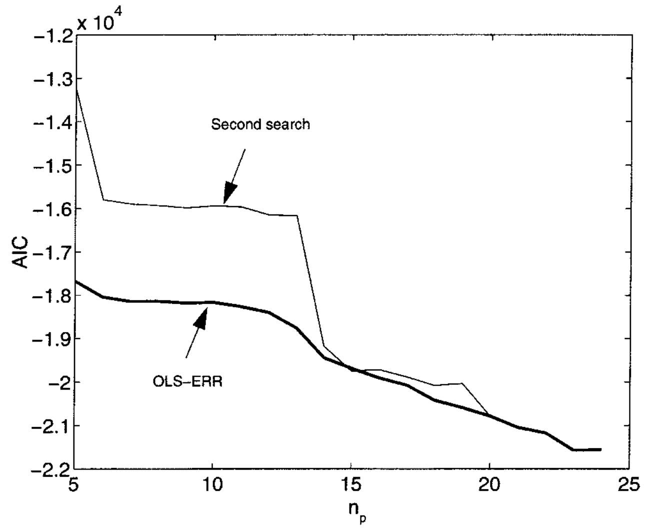


Fig. 6. AIC criterion for estimated models with increasing number of terms. The upper curve shows the values of AIC for estimated models using the second search procedure. Whereas the lower curve shows the values of AIC for estimated models using OLS-ERR procedure.

using the OLS-ERR procedure, a few models are found to be able reproduce the original dynamic invariants of the systema. Despite the initial success of the OLS-ERR procedure, the valid models proved to be unstable when they were iterated for a long time. The use of a second term search lead to better results, that is, stable and dynamically valid models for Rossler equations. Details of the results are available from the authors.

V. CONCLUSION

In this work, an alternative method for solving the model structure selection problem has been proposed and analyzed.

A procedure which combines the advantages of the OLS-ERR and exhaustive-like algorithms has been investigated. Such a procedure can successfully identify not only a single valid model but a family of models. Such family of models were shown to preserve the desired dynamic characteristics of the original system. It is believed that the results obtained in this paper will help the user of identification techniques, specially when complex nonlinear phenomena such as chaos is of interest. An example has been provided to demonstrate the usefulness of this approach.

As a by-product, a new but simple procedure to determine to maximum lag (embedding dimension) has been devised. This procedure has been used to find the maximum lag for the Chua's circuit.

It has also been demonstrated that the so-called Information criteria which are extensively used to determine the trade-off between complexity and fitness accuracy can be strongly affected by the inclusion of noise terms. Alternative criteria using dynamical quantities such as the Lyapunov dimension and spectrum have been proposed for selecting valid models. The simulated example shows the efficacy of such a choice.

REFERENCES

[1] R. R. Hocking, "The analysis and selection of variables in linear regression," *Biometrika*, vol. 32, pp. 1–49, Mar. 1976.

[2] J. W. Gorman and R. J. Tomam, "Selection of variables for fitting equations to data," *Technometrics*, vol. 8, pp. 27–51, 1966.

[3] J. C. Principe, A. Rathie, and J. M. Kuo, "Prediction of chaotic time series with neural networks and the issue of dynamic modeling," *Int. J. Bif. Chaos*, vol. 2, no. 4, pp. 989–996, 1992.

[4] M. J. Garside, "The best subset in multiple regression analysis," *Appl. Statistician*, vol. 14, pp. 196–200, 1965.

[5] M. Schatzoff, R. Tsao, and S. Fienberg, "Efficient calculation of all possible regressions," *Technometrics*, vol. 10, no. 4, pp. 769–779, Nov. 1968.

[6] G. M. Furnival and R. W. Wilson, "Regression by leaps and bounds," *Technometrics*, vol. 16, pp. 499–511, 1974.

[7] D. Edwards and T. Havranek, "A fast model selection procedure for large families of model," *J. Amer. Statist. Assoc.*, vol. 82, pp. 205–213, 1987.

[8] A. J. Miller, *Subset Selection in Regression*. London, U.K.: Chapman and Hall, 1990.

[9] A. Kudo and T. Tarumi, "An algorithm related to all possible regression and discriminant analysis," *J. Japan Statist. Soc.*, vol. 4, pp. 47–56, 1974.

[10] S. C. Narula and J. F. Wellington, "Selection of variables in linear regression using the sum of weighted absolute errors criterion," *Technometrics*, vol. 21, pp. 299–306, 1979.

[11] R. R. Hocking, "Developments in linear regression methodology: 1959–1982 (with discussion)," *Technometrics*, vol. 25, no. 3, pp. 219–249, Aug. 1983.

[12] M. A. Efroymson, "Multiple regression analysis," in *Mathematical Methods for Digital Computers*, A. Ralston and H. S. Wilf, Eds. New York: Wiley, 1960.

[13] S. A. Billings, S. Chen, and M. J. Korenberg, "Identification of MIMO nonlinear systems using a forward-regression orthogonal estimator," *Int. J. Contr.*, vol. 49, no. 6, pp. 2157–2189, 1989.

[14] D. E. O. Boyce, A. Farhi, and R. Weischedel, *Optimal Subset Selection, Multiple Regression, Interdependence and Optimal Network Algorithms*. New York: Springer-Verlag, 1974.

[15] I. J. Leontaritis and S. A. Billings, "Input–output parametric models for nonlinear systems—Part I: Deterministic nonlinear systems," *Int. J. Contr.*, vol. 41, no. 2, pp. 303–328, 1985.

[16] S. A. Billings and I. J. Leontaritis, "Identification of nonlinear systems using parameter estimation techniques," in *Proc. Inst. Elect. Eng. Conf. Contr. Applicat.*, Warwick, U.K, 1981, pp. 183–187.

[17] J. Hammer, "Nonlinear systems, stability and rationality," *Int. J. Contr.*, vol. 40, pp. 1–35, 1984.

[18] S. A. Billings, S. Chen, and R. J. Backhouse, "Identification of linear and nonlinear models of a turbocharged automotive diesel engine," *Mech. Syst. Signal Process.*, vol. 3, no. 2, pp. 123–142, 1989.

[19] S. Chen and S. A. Billings, "Representations of nonlinear systems: The NARMAX model," *Int. J. Contr.*, vol. 49, no. 3, pp. 1013–1032, 1989.

[20] S. A. Billings, M. Korenberg, and S. Chen, "Identification of nonlinear output-affine systems using an orthogonal least-squares algorithm," *Int. J. Syst. Sci.*, vol. 19, no. 8, pp. 1559–1568, 1988.

[21] S. A. Billings and W. S. F. Voon, "Least squares parameter estimation algorithms for nonlinear systems," *Int. J. Syst. Sci.*, vol. 15, no. 6, pp. 601–615, 1984.

[22] J. C. Peyton-Jones and S. A. Billings, "Recursive algorithms for computing the frequency response of a class of nonlinear difference equation models," *Int. J. Contr.*, vol. 50, pp. 1925–1940, 1989.

[23] L. A. Aguirre and S. A. Billings, "Improved structure selection for nonlinear models based on term clustering," *Int. J. Contr.*, vol. 62, pp. 569–587, 1995.

[24] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcation of Vector Fields*. New York: Springer-Verlag, 1983.

[25] E. M. A. M. Mendes and S. A. Billings, "On identifying global nonlinear discrete models from chaotic data," *Int. J. Bif. Chaos*, vol. 7, no. 11, pp. 2593–2602, Nov. 1997.

[26] T. Söderström and P. Stoica, *System Identification*. Englewood Cliffs, NJ: Prentice-Hall, 1989.

[27] S. A. Billings and W. S. F. Voon, "Correlation based model validity tests for nonlinear models," *Int. J. Contr.*, vol. 44, no. 1, pp. 235–244, 1986.

[28] S. A. Billings and Q. H. Tao, "Model validity tests for nonlinear signal processing applications," *Int. J. Contr.*, vol. 54, no. 1, pp. 157–194, 1991.

[29] S. A. Billings and Q. M. Zhu, "Nonlinear model validation using correlation tests," *Int. J. Contr.*, vol. 60, pp. 1107–1120, 1994.

[30] L. A. Aguirre and S. A. Billings, "Dynamical effects of overparameterization in nonlinear models," *Physica D*, vol. 80, no. 1 & 2, pp. 26–40, 1995.

[31] L. O. Chua, "Chua's circuit: An overview ten years later," *J. Circuits, Syst. Comput.*, vol. 4, no. 2, pp. 117–159, 1994.

[32] E. M. A. M. Mendes and S. A. Billings, "Discretizing a nonlinear analytic system: Consequences on the model structure problem," FUNREI Tech Rep., 2001.

[33] G. Rowland and J. C. Sprott, "Extraction of dynamical equations from chaotic data," *Physica D*, vol. 58, pp. 251–259, 1992.

[34] E. N. Lorenz, "Deterministic nonperiodic flow," *J. Atmospheric Sci.*, vol. 20, pp. 131–141, Mar. 1963.

[35] O. E. Rossler, "An equation for continuous chaos," *Phys. Lett. A*, vol. 57, no. 5, pp. 397–398, July 1976.

[36] J. Glover and A. I. Mees, "Reconstructing the dynamics of Chua's circuit," *J. Circuits, Syst. Comput.*, vol. 3, no. 1, pp. 201–214, 1993.

[37] E. M. A. M. Mendes and S. A. Billings, "On overparameterization of global nonlinear discrete models," *Int. J. Bif. Chaos*, vol. 8, no. 3, pp. 535–556, Mar. 1998.

[38] F. Takens, "Detecting strange attractors in fluid turbulance," in *Lecture Notes in Mathematics*, R. A. Rand and L. S. Young, Eds. New York: Springer-Verlag, 1981, vol. 898, pp. 366–381.

[39] D. S. Broomhead, R. Jones, and G. P. King, "Topological dimension and local coordinates for time series data," *J. Phys. A: Math. Gen.*, vol. 20, pp. L563–L569, 1987.

[40] F. C. Moon, *Chaotic Vibrations—An Introduction for Applied Scientists and Engineers*. New York: Wiley, 1987.

[41] H. D. I. Abarbanel and M. B. Kennel, "Local false nearest neighbors and dynamical dimensions from observer chaotic data," *Phys. Rev. E*, vol. 47, no. 5, pp. 3057–3068, 1993.

[42] R. Savit and M. L. Green, "Time series and dependent variables," *Physica D*, vol. 50, pp. 95–116, 1991.

[43] Y. Ueda, "Random phenomena resulting from nonlinearity in the system described by Duffing's equation," *Int. J. Non-Linear Mech.*, vol. 20, no. 5/6, pp. 481–491, 1985.

[44] P. J. Holmes, "A nonlinear oscillator with a strange attractor," *Phil. Trans. Roy. Soc. London A*, vol. 292, pp. 419–448, 1979.

[45] L. A. Aguirre and S. A. Billings, "Discrete reconstruction of strange attractors in Chua's circuit," *Int. J. Bif. Chaos*, vol. 4, no. 4, pp. 853–864, 1994.

[46] J. Theiler and S. Eubank, "Don't bleach chaotic data," *Chaos*, vol. 4, Dec. 1993.

[47] M. Casdagli, S. Eubank, J. D. Farmer, and J. Gibson, "State space reconstruction in the presence of noise," *Physica D*, vol. 51, pp. 52–98, 1991.

[48] M. Korenberg, "Orthogonal identification of nonlinear difference equation models," in *Proc. 28th Midwest Symp. Circuits Syst.*, Louisville, KY, 1985, pp. 90–95.

[49] ——, "Fast orthogonal identification of nonlinear difference equation and functional expansion models," in *Proc. 30th Midwest Symp. Circuits Syst.*, Aug. 1987, pp. 270–276.

[50] M. Korenberg, S. A. Billings, Y. P. Liu, and P. J. McIlroy, "Orthogonal parameter estimation algorithm for nonlinear stochastic systems," *Int. J. Contr.*, vol. 48, no. 1, pp. 193–210, 1988.

[51] M. Pottmann, H. Unbehauen, and D. E. Seborg, "Application of a general multi-model approach for identification of a highly nonlinear system—A case study," *Int. J. Contr.*, vol. 57, no. 1, pp. 197–120, Jan. 1993.

[52] R. R. Hocking and R. N. Leslie, "Selection of the best subset in regression analysis," *Technometrics*, vol. 9, pp. 531–540, 1967.

[53] M. Stone, "Cross-validatory choice and assessment of statistical predictions," *J. Roy. Statist. Soc. Ser. B*, vol. 36, pp. 111–147, 1974.

[54] L. Lennard and G. Torkel, "Modeling of dynamic systems," in *Information and System Sciences Series*. Englewood Cliffs, NJ: Prentice-Hall, 1994.

[55] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Automat. Contr.*, vol. AC-19, pp. 716–723, 1974.

[56] E. J. Hannan and B. G. Quinn, "The determination of the order of an autoregression," *J. R. Statist. Soc. B*, vol. 41, no. 2, pp. 190–195, 1979.

[57] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, pp. 416–464, 1978.

[58] J. Rissanen, "Stochastic complexity and modeling," *Ann. Statist.*, vol. 14, no. 3, pp. 1080–1100, 1986.

[59] H. Akaike, "On entropy maximization principle," in *Applications of Statistics*, P. R. Krishnaiah, Ed. Amsterdam, The Netherlands: North-Holland, 1977, pp. 27–41.

[60] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.

[61] M. Stone, "An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion," *J. R. Statist. Soc. Ser. B*, vol. 39, pp. 44–47, 1977.

[62] C. L. Mallows, "Some comments on c_p," *Technometrics*, vol. 15, pp. 661–675, 1973.

[63] D. M. Allen, "Mean square error of prediction as a criterion for selection variables," *Technometrics*, vol. 13, pp. 469–475, 1971.

[64] J. P. Crutchfield and B. S. McNamara, "Equations of motion from a data series," *Complex Syst.*, vol. 1, pp. 417–452, 1987.

[65] J. G. Gooijer, B. Abraham, A. Gould, and L. Robinson, "Methods for determining the order of an autoregressive-moving average process: A survey," *Int. Statist. Rev.*, vol. 53, no. 3, pp. 301–329, 1985.

[66] C. Z. Wei, "On predictive least squares principles," *Ann. Statist.*, vol. 20, no. 1, pp. 1–42, 1992.

[67] M. Kortmann and H. Unbehauen, "Two algorithms for model structure determination of nonlinear dynamic systems with applications to industrial process," in *Proc. IFAC Symp. Identification Syst. Parameter Estimation*, Beijing, China, 1988, pp. 649–656.

[68] A. I. Mees, "Parsimonious dynamical reconstruction," *Int. J. Bif. Chaos*, vol. 3, no. 3, pp. 669–675, 1993.

Eduardo M. A. M. Mendes was born in Belo Horizonte, Brazil, in 1964. He received the B.Eng. degree in electrical engineering (with class honors) and the M.Sc. degree from Federal University of Minas Gerais, Belo Horizonte, in 1988 and 1991, respectively. He received the Ph.D. degree in control systems engineering from the University of Sheffield, Sheffield, U.K., in 1995.

He was appointed to his current position of Associate Professor, Departamento de Eletricidade, Fundação de Ensino Superior de São João Del Rei, São João Del Rei, Brazil, in 1996. His research interests include system identification for nonlinear systems, narmax methods, model validation, prediction, spectral analysis and chaos.

Steve A. Billings was born in Staffordshire, U.K., in 1951. He received the B.Eng. degree in electrical engineering (first class honors) from the University of Liverpool, Liverpool, U.K., the Ph.D. degree in control systems engineering from the University of Sheffield, Sheffield, U.K., and the D.Eng. degree from the University of Liverpool, in 1972, 1976, and 1990, respectively.

He was appointed to his current position of Professor, Department of Automatic Control and Systems Engineering, University of Sheffield, in 1990 and leads the Signal Processing and Complex Systems Research Group. His research interests include system identification and information processing for nonlinear systems, narmax methods, model validation, prediction, spectral analysis, adaptive systems, nonlinear systems analysis and design, neural networks, machine vision, spatio-temporal systems and brain imaging.

Dr. Billings is a Chartered Engineer [C.Eng.], Chartered Mathematician [C.Math.], Fellow of the IEE [U.K.], and Fellow of the Institute of Mathematics and its Applications.